CLAIMS

1. A method comprising the step of:

cleaning, by operations of a computer system, a set of text documents to minimize violations of a predetermined set of Hypertext Information Retrieval rules.

2. The method of claim 1, wherein the set of text documents comprises a collection of HTML pages.

3. The method of claim 1, wherein the cleaning step comprises the steps of:

decomposing each page of the set of text documents into one or more pagelets;

identifying all pagelets belonging to templates; and

eliminating the template pagelets from a data set.

4. The method of claim 3, wherein the decomposing step comprises the steps of:

parsing each text document into a parse tree that comprises at least one node;

traversing the at least one node of the tree;

determining if one of the at least one node comprises a pagelet; and

outputting a representation corresponding to the one of the at least one node if it comprises a pagelet.

5. The method of claim 4, wherein the determining step comprises the steps of:

verifying the node is of a type belonging to a predetermined class of eligible types;

verifying the node contains at least a predetermined number of hyperlinks; and

verifying none of the node's children are pagelets.

6. The method of claim 5, wherein the predetermined class of eligible types comprises at least one of tables, lists, paragraphs, image maps, headers, table rows, table cells, list items, selection bars, and frames.

7. The method of claim 3, wherein the step of identifying all pagelets belonging to templates comprises the steps of:

calculating a shingle value for each page and for each pagelet in the set of documents;

eliminating identical pagelets belonging to duplicate pages;

sorting the pagelets by their shingle value into clusters;

enumerating the clusters; and

outputting a representation corresponding to the pagelets belonging to each cluster.

8. The method of claim 3, wherein the step of identifying pagelets belonging to templates comprises the steps of:

calculating a shingle value for each page and for each pagelet in the document set;

sorting the pagelets by their shingle value into clusters;

selecting all clusters of size greater than 1;

finding for each cluster all hyperlinks between pages owning pagelets in that cluster;

finding for each cluster all undirected connected components of a graph induced by the pages owning pagelets in that cluster; and

outputting a representation corresponding to the components of size greater than 1.

9. A system comprising:

a user interface;

a user interface/event manager communicatively coupled to the user interface;

a generic data gathering application;

a generic information retrieval application, communicatively coupled to the user interface/event manger; and

a data cleaning application for:

decomposing each page of a set of text documents into one or more pagelets;

identifying all pagelets belonging to templates; and

eliminating the template pagelets from a data set,

communicatively coupled to the generic data gathering application and to the generic information retrieval application.

10. The system of claim 9, further comprising:

a pagelet identifier, communicatively coupled to the data cleaning application;

a hypertext parser, communicatively coupled to the pagelet identifier;

a template identifier, communicatively coupled to the data cleaning application;

a Breadth First Search (BFS) algorithm, communicatively coupled to the template identifier; and

a shingle calculator, communicatively coupled to the data cleaning application.

11. An apparatus comprising:

a user interface;

a user interface/event manager communicatively coupled to the user interface;

a generic data gathering application;

a generic information retrieval application, communicatively coupled to the user interface/event manger; and

a data cleaning application, for:

decomposing each page of the set of text documents into one or more pagelets;

identifying all pagelets belonging to templates; and

eliminating the template pagelets from a data set,

communicatively coupled to the generic data gathering application and to the generic information retrieval application.

12. The apparatus of claim 11, further comprising:

a pagelet identifier, communicatively coupled to the data cleaning application;

a hypertext parser, communicatively coupled to the pagelet identifier;

a template identifier, communicatively coupled to the data cleaning application;

a BFS algorithm, communicatively coupled to the template identifier; and

a shingle calculator, communicatively coupled to the data cleaning application.

13. A computer readable medium including computer instructions for driving a user interface, the computer instructions comprising instructions for:

cleaning, by operations of a computer system, a set of text documents to minimize violations of a predetermined set of Hypertext Information Retrieval rules.

14. The computer readable medium of claim 13, wherein the set of text documents comprises a collection of HTML pages.

15. The computer readable medium of claim 13, wherein the cleaning step comprises the steps of:

decomposing each page of the set of text documents into one or more pagelets;

identifying any pagelets belonging to templates; and

eliminating the template pagelets from a data set.

16. The computer readable medium of claim 15, wherein the decomposing step comprises the steps of:

parsing each text document into a parse tree that comprises at least one node;

traversing the at least one node of the tree;

determining if one of the at least one node comprises a pagelet; and

outputting a representation corresponding to the one of the at least one node if it comprises a pagelet.

17. The computer readable medium of claim 16, wherein the determining step comprises the steps of:

verifying the node is of a type belonging to a predetermined class of eligible types;

verifying the node contains at least a predetermined number of hyperlinks; and

verifying none of the node's children are pagelets.

18. The computer readable medium of claim 17, wherein the predetermined class of eligible types comprises at least one of tables, lists, paragraphs, image maps, headers, table rows, table cells, list items, selection bars, and frames.

19. The computer readable medium of claim 15, wherein the step of identifying pagelets belonging to templates comprises the steps of:

calculating a shingle value for each page and for each pagelet in the set of documents;

eliminating identical pagelets belonging to duplicate pages;

sorting the pagelets by their shingle value into clusters;

enumerating the clusters; and

outputting a representation corresponding to the pagelets belonging to each cluster.

20. The computer readable medium of claim 15, wherein the step of identifying pagelets belonging to templates comprises the steps of:

calculating a shingle value for each page and for each pagelet in the document set;

sorting the pagelets by their shingle value into clusters;

selecting all clusters of size greater than 1;

finding for each cluster all hyperlinks between pages owning pagelets in that cluster;

finding for each cluster all undirected connected components of a graph induced by the pages owning pagelets in that cluster; and

outputting a representation corresponding to the components of size greater than 1.